European Journal of Obstetrics and Gynecology 302 (2024) 238–241

Contents lists available at ScienceDirect

European Journal of Obstetrics & Gynecology and Reproductive Biology

journal homepage: www.journals.elsevier.com/european-journal-of-obstetrics-and-gynecology-and-reproductive-biology

Full length article

AI in obstetrics: Evaluating residents' capabilities and interaction strategies with ChatGPT

David Desseauve [a,b,*], Raphael Lescar [c], Benoit de la Fourniere [c], Pierre-François Ceccaldi [d,e], Mikhail Dziadzko [f,g]

[a] Department of Women–Mother–Child, Gynaecology and Obstetrics Unit, Lausanne University Hospital, Lausanne, Switzerland
[b] Department of Women–Mother–Child, Gynaecology and Obstetrics Unit, Grenoble Alpes, University Hospital, Grenoble, France
[c] Department of Obstetrics and Gynaecology, Hôpital de la Croix-Rousse, Hospices civils de Lyon, Lyon, France
[d] Department of Obstetrics, Gynaecology and Reproductive Medicine, Foch Hospital, Suresnes, France
[e] Innovative Dental Materials and Interfaces Research Unit (UR 4462), Faculty of Health, University of Paris, Paris, France
[f] Department of Anaesthesiology, Hôpital de la Croix-Rousse, Hospices civils de Lyon, Lyon, France
[g] RESHAPE UMR 1290 INSERM, Université Lyon 1, Lyon, France

ARTICLE INFO

Keywords:
Artificial Intelligence
Obstetrics
Medical Education
Prompt Engineering

ABSTRACT

In line with the digital transformation trend in medical training, students may resort to artificial intelligence (AI) for learning. This study assessed the interaction between obstetrics residents and ChatGPT during clinically oriented summative evaluations related to acute hepatic steatosis of pregnancy, and their self-reported competencies in information technology (IT) and AI.

The participants in this semi-qualitative observational study were 14 obstetrics residents from two university hospitals. Students' queries were categorized into three distinct types: third-party enquiries; search-engine-style queries; and GPT-centric prompts. Responses were compared against a standardized answer produced by ChatGPT with a Delphi-developed expert prompt. Data analysis employed descriptive statistics and correlation analysis to explore the relationship between AI/IT skills and response accuracy.

The study participants showed moderate IT proficiency but low AI proficiency. Interaction with ChatGPT regarding clinical signs of acute hepatic steatosis gravidarum revealed a preference for third-party questioning, resulting in only 21% accurate responses due to misinterpretation of medical acronyms. No correlation was found between AI response accuracy and the residents' self-assessed IT or AI skills, with most expressing dissatisfaction with their AI training. This study underlines the discrepancy between perceived and actual AI proficiency, highlighted by clinically inaccurate yet plausible AI responses – a manifestation of the 'stochastic parrot' phenomenon.

These findings advocate for the inclusion of structured AI literacy programmes in medical education, focusing on prompt engineering. These academic skills are essential to exploit AI's potential in obstetrics and gynaecology. The ultimate aim is to optimize patient care in AI-augmented health care, and prevent misleading and unsafe knowledge acquisition.

Introduction

In recent decades, digital technologies have transformed medical education. Human–computer interactions have become essential in medical training, including various activities such as information retrieval, knowledge assessment, and simulation [1]. The emergence of artificial intelligence (AI) holds the potential to improve medical education further, offering personalized and interactive learning experiences, enhancing diagnostic accuracy, and providing support for data-driven decision-making [1–3].

Applications of AI in obstetrics and gynaecology (OG) training have shown diverse results. Initially, AI achieved an initial success rate of 30 %, which was increased to 70 % when used to answer questions on various topics usually asked to first-year medical students. This increase was obtained after applying an iterative querying technique called 'prompting' [4,5]. The precision of AI-generated answers is contingent

* Corresponding author at: Department of Women–Mother–Child, Lausanne University Hospital, 1011 Lausanne, Switzerland.
E-mail address: david.desseauve@chuv.ch (D. Desseauve).

https://doi.org/10.1016/j.ejogrb.2024.09.008
Received 22 December 2023; Received in revised form 1 June 2024; Accepted 6 September 2024
Available online 14 September 2024

on the methodology employed in questioning.

The emergence of 'prompt engineering' as a skill for leveraging the capabilities of large language models (LLMs) such as ChatGPT has not yet been integrated systematically into medical training [6,7]. Although not adopted universally, several recommendations and techniques in medical prompt engineering have been published, highlighting the significance of cultivating this skill within students' curricula [8].

This descriptive study was designed to understand how obstetrics and gynaecology residents interact spontaneously with AI in the context of a summative training session. The main objective was to study the accuracy of GPTs' answers depending on the methods used by obstetrics and gynaecology residents to question the AI. The secondary objective was to examine the correlation between self-assessed IT proficiency, AI answers, AI skills and the participants' postgraduate year (PGY).

Methods

Study design and setting

This semi-qualitative observational study was conducted in a controlled academic environment to explore how obstetrics residents interact and seek answers from ChatGPT [9], a renowned LLM designed by OpenAI [10]. The sessions took place in two university hospital centres.

Participants

The study cohort comprised obstetrics residents, chosen from two hospitals to ensure diversity in training backgrounds. No ethical approval was required; written consent was obtained for voluntary participation.

AI interaction

Residents were asked to answer the question (in French): 'What are the clinical and paraclinical signs of acute fatty liver of pregnancy (AFLP)?'. AFLP is a rare but severe liver condition that can occur during pregnancy, and can be life-threatening for both the mother and the baby. With prompt diagnosis and treatment, most women with AFLP recover fully after delivery. Every resident had access to their personal computer to facilitate interaction with ChatGPT 3.5. The time to obtain the answer was arbitrarily fixed to 15 min.

Expert prompt and standard answer

Using a Delphi method, the educational team (DD, MD, BdLF) created an expert prompt using 'prompt engineering methods'. This process involves creating prompts by iterative refining and revising the initial queries to elicit accurate responses. For instance, a prompt might instruct: 'Imagine you're a PGY 3 medical student specializing in obstetrics. Generate three distinct queries for ChatGPT to obtain the following responses: [desired response in brackets]'. The authors collaboratively reviewed and edited the prompts until consensus was reached on the final version.

This prompt was used to obtain the result, and it was submitted to three experts in the management of AFLP (GD, CH, PFC) from two medical schools for evaluation and validation (see online supplementary material). The validated result from the constructed prompt was used as a standard comparator for the students' results obtained from AI.

Data collection procedure

Before the main task, the participants' PGY, a self-assessment of their IT and AI proficiency (11-item scale, where 0 represents no proficiency and 10 represents maximal proficiency), and previous experience of AI use (yes/no) were collected. The queries used with ChatGPT were recorded and yielded answers. At the end of the study, participants' satisfaction with AI interaction was recorded (11-item scale, where 0 represents no satisfaction and 10 represents maximal satisfaction).

Analytical approach

All queries were analysed and categorized semantically into the following types:

- Third-party question: questions were structured as if asking a third person or expert.
- Search-engine style: questions were formatted similarly to what one might input into a search engine.
- GPT-centric prompt: questions were tailored explicitly to the known interaction style of GPT-like models [8].

All responses were compared with the standard comparator, and evaluated using a French grading system out of 20 points for their medical accuracy and relevance to the original clinical question. Based on a scale of 0–20 points, this grading system is commonly used in French educational institutions. Scores < 10 denote failure, with 0 representing complete failure. Scores > 10 indicate varying degrees of success, with a score of 20 signifying exceptional performance. The final score was converted to an accuracy percentage.

Descriptive statistics and a non-parametric Spearman's rho correlation analysis were used to evaluate the relationships between self-reported AI and IT skills and the accuracy of responses obtained from ChatGPT using students' queries. This exploratory study did not use a sample size calculation.

Data are presented as median [interquartile range (IQR)] or frequency (percentage), as appropriate. $p < 0.05$ was considered to indicate significance.

Results

Demographic and background information

The study included 14 obstetrics residents from two university hospitals. Half of them (7/14) were at the end of their residency (more than eight semesters), while another 43 % (6/14) were in the middle of their residency (three to eight semesters). Only one resident was at the beginning of their residency. Only 36 % of the residents (5/14) had previous experience with LLMs such as GPT or BARD.

The median levels of self-assessed IT proficiency and AI proficiency were 6 [IQR 5.75–7] and 2 [IQR 1–4.25], respectively. Eleven (93 %) students rated their IT proficiency higher than 5 out of 10, and only three (21 %) students gave such a rating for their AI skills.

Interaction with ChatGPT

The primary task for participants was to enquire about the clinical and paraclinical signs of acute hepatic steatosis gravidarum (commonly abbreviated as 'SHAG' in French). The majority (64 %, 9/14) chose the third-party question method, mainly copying the same question given by the examiner. Only one resident employed a prompt-like technique; the remaining 29 % (4/14) used a search-engine-style approach.

From the entire cohort, only three (21 %) of the GPT answers corresponded to the expected response, with accuracy of 30 % (one answer), 40 % (one answer) and 90 % (one answer). Predominantly, inaccuracies originated from a misinterpretation of the 'SHAG' acronym by AI, resulting in unrelated answers.

Accuracy of responses and relationships with IT/AI skills

No correlation was found between the accuracy of the AI response and self-assessed IT or AI skills, nor for the PGY level. Strong correlation

was found between self-assessed IT and AI skills (Table 1). In an open question, 86 % (12/14) of residents expressed concerns about their AI training, stating that their current education needed to be equipped sufficiently for AI utilization. Only two (14 %) participants were satisfied with their level of AI knowledge.

Discussion

This study revealed challenges in AI interaction among obstetrics and gynaecology residents, with low accuracy rates in ChatGPT-generated answers. No correlation was found between response accuracy and residents' self-assessed IT or AI skills. Only a minority of participants used the prompting method.

One of the most significant advances in AI chatbots is the claimed ability to 'understand' human language conversations within context. However, the accuracy of an AI response, even with precise contextual information, can be outweighed by the 'stochastic parrot' [11]. Iterative prompting and clarification of questions by providing additional background information and context may enhance the accuracy of responses. This was observed in this study with the efficacy of questioning when using a prompt-like approach. Additionally, the limited number of correct responses to the expert-level question aligns with prior research demonstrating diminished performance of LLMs as the complexity of questions increases [12].

The difference in the results between responders' self-assessed IT and AI competencies is significant. Although correlated, conventional digital knowledge does not equate to AI competency because the latter is highly dependent on specific skills such as prompting. This study demonstrated that many students lack understanding regarding AI interactions and their implications [15].

Based on these findings, a primary concern highlighted by this study regarding the integration of AI into medical practice or education is the risk of yielding inaccurate results, particularly in scenarios involving a combination of summative evaluation (assigning a score to students' answers) and formative training (constructing new knowledge). This limit is also well known in using usual internet navigators with integrated web-search machines to answer questions in other medical fields [16–18]. Nevertheless, the performance of LLMs and the apparent quality of the answers may mask potential inaccuracies, possibly leading students or physicians to dramatic errors if they are applied directly to patient care without expert oversight. Applying evidence-based medicine mitigates these risks, particularly in the management of AFLP. Adhering to clinical guidelines and peer-reviewed research ensures that decisions are based on the best-available evidence, reducing the potential for errors from LLM-generated recommendations.

One potential solution is to promote the incorporation of prompt engineering in medical education, in parallel with the exponential AI implementation in medical practice.

Broad questions are likely to produce incorrect results. This underscores the importance of possessing AI prompt-building skills, and understanding how to construct medical or professional questions effectively to provide a structure for the anticipated answer. An alternative approach to developing effective prompt-engineering skills involves the reverse prompting technique, in which the LLM generates a prompt in a preconditioned manner based on an arbitrarily correct example of an answer. Subsequently, a medical student is invited to formulate a question for the LLM, aiming to elicit a response closely aligned with the initial answer, though not necessarily replicating the prompt originally generated by the LLM. Multiple iterations may be required to achieve the desired level of accuracy in prompt construction.

Prompt engineering, associating the art of building precise textual prompts, is essential in interacting effectively with chatbots and generative AI tools to bring out desired outputs across various digital media formats [19]. The quality of these prompts is crucial, as AI models can improve their accuracy through iterative learning from user-provided data [20]. This emerging discipline is gradually evolving into a distinct skill set in the technology and corporate sectors. In medical training and education, understanding generative AI principles is critical to produce outcomes that support teaching, learning and assessment.

In this way, Acar recently introduced a concept known as the Problem, AI, Interaction, Reflection (PAIR) framework [21]. This framework offers a comprehensive roadmap for leveraging generative AI tools, mainly focusing on prompt development and its application. Widely recognized within academic circles, the PAIR framework serves as a cornerstone in prompt engineering, focusing on the critical role of problem formulation and understanding. The depth and precision of comprehension of this problem are pivotal, as they dictate the choice of prompts employed, the ensuing responses generated by the AI tool, and, ultimately, the effectiveness of the entire process. Such a condition requires human expertise.

This underscores the importance of integrating structured AI utilization training modules into medical curricula, ensuring that the forthcoming generation of practitioners is both AI-aware and AI-competent [13]. As emphasized by the Beijing Convention on AI and Education [14], educational programmes are responsible for investing in and instilling these skills.

This study has a few limitations. The modest sample size and the study's confinement to two academic settings may have restricted the generalizability of the findings. However, it is unlikely that more advanced AI skills would be expected among students in other French medical universities because no LLM engineering skills are currently integrated into medical curricula. Furthermore, while this study assessed the immediate interactions with AI, no semantic analysis or incremental engineering prompt analysis (prompt-based incremental learning) was performed. The correlation analysis was performed on 14 pairs of responses. Although this number is modest and does not produce a critical side effect, the Spearman rank-order correlation analysis is valid for such a small number of observations [22].

A longitudinal analysis in further studies would provide insights into the learning curve of mastering AI tools.

Table 1
Relationships between the accuracy of obtained artificial intelligence (AI) response, self-rated information technology (IT) and AI experience, and post-graduate year (PGY).

	Accuracy	IT level	AI level	PGY
Accuracy	1	0.1216; $p = 0.68$	0.2914; $p = 0.31$	0.5744; $p = 0.10$
IT level		1	0.8557; $p < 0.0001$	0.1650; $p = 0.57$
AI level			1	0.4503; $p = 0.11$
PGY				1

Numbers are Spearman's rho and *p*-values.

Conclusion

This study underscores a significant finding: non-AI-trained obstetric gynaecology residents exhibit a heightened propensity to employ 'third-party-like questions' when interfacing with LLMs such as ChatGPT. This worrisome trend results in the dissemination of incorrect and potentially detrimental information within the context of medical education. It is evident that prompt engineering, a burgeoning facet of human–computer interaction, demands immediate integration into medical curricula. The precise role of AI as a facilitator in medical training remains to be defined comprehensively, leaving an imperative void in the field that requires scholarly attention and resolution.

Funding

None.

CRediT authorship contribution statement

David Desseauve: Writing – review & editing, Writing – original draft, Validation, Supervision, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Raphael Lescar:** Writing – original draft, Investigation, Formal analysis, Data curation. **Benoit de la Fourniere:** Writing – review & editing, Data curation. **Pierre François Ceccaldi:** Writing – review & editing, Project administration, Methodology, Investigation, Formal analysis, Conceptualization. **Mikhail Dziadzko:** Writing – review & editing, Supervision, Methodology, Formal analysis.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors wish to thank the two department heads, Pr Cyril Huissoud (CH) and Pr Gil Dubernard (GD), for their support and guidance in enabling the successful completion of this study.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ejogrb.2024.09.008.

References

[1] Dave M, Patel N. Artificial intelligence in healthcare and education. Br Dent J 2023;234:761–4.
[2] Nagi F, Salih R, Alzubaidi M, et al. Applications of artificial intelligence (AI) in medical education: a scoping review. In: Mantas J, Gallos P, Zoulias E, editors. Studies in health technology and informatics. IOS Press; 2023. Available at: https://ebooks.iospress.nl/doi/10.3233/SHTI230581 (last accessed 17 October 2023).
[3] Alowais SA, Alghamdi SS, Alsuhebany N, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ 2023;23:689.
[4] Li SW, Kemp MW, Logan SJS, et al. ChatGPT outscored human candidates in a virtual objective structured clinical examination in obstetrics and gynecology. Am J Obstet Gynecol 2023;229(172). 172.e1–12.
[5] Guerra GA, Hofmann H, Sobhani S, et al. GPT-4 artificial intelligence model outperforms ChatGPT, medical students, and neurosurgery residents on neurosurgery written board-like questions. World Neurosurg 2023;179:e160–5.
[6] Wang L, Bi W, Zhao S, Ma Y, Lv L, Meng C, et al. Investigating the impact of prompt engineering on the performance of large language models for standardizing obstetric diagnosis text: comparative study. JMIR Form Res 2024;8:e53216.
[7] O'Connor S, Peltonen LM, Topaz M, et al. Prompt engineering when using generative AI in nursing education. Nurse Educ Pract 2024;74:103825.
[8] Meskó B. Prompt engineering as an important emerging skill for medical professionals: tutorial. J Med Internet Res 2023;25:e50638.
[9] ChatGPT. Available at: https://chat.openai.com (last accessed 18 September 2023).
[10] OpenAI. Available at: https://openai.com/ (last accessed 18 September 2023).
[11] Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: can language models be too big?. In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. ACM; 2021. p. 610–23. Available at: https://dl.acm.org/doi/10.1145/3442188.3445922 (last accessed 17 October 2023).
[12] Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023;2:e0000198.
[13] Nune A, Iyengar Karthikeyan P, Manzo C, Barman B, Botchu R. Chat generative pre-trained transformer (ChatGPT): potential implications for rheumatology practice. Rheumatol Int 2023;43:1379–80.
[14] Grünebaum A, Chervenak J, Pollet SL, Katz A, Chervenak FA. The exciting potential for ChatGPT in obstetrics and gynecology. Am J Obstet Gynecol 2023; 228:696–705.
[15] Pinto Dos Santos D, Giese D, Brodehl S, et al. Medical students' attitude towards artificial intelligence: a multicentre survey. Eur Radiol 2019;29:1640–6.
[16] Kothari M, Moolani S. Reliability of 'Google' for obtaining medical information. Indian J Ophthalmol 2015;63:267–9.
[17] Al-Bahrani A, Plusa S. The quality of patient-orientated internet information on colorectal cancer. Colorectal Dis 2004;6:323–6.
[18] Bristowe K, Siassakos D, Hambly H, et al. Teamwork for clinical emergencies: interprofessional focus group analysis and triangulation with simulation. Qual Health Res 2012;22:1383–94.
[19] Prompt engineering guide. Available at: https://www.promptingguide.ai/fr/techniques/cot (last accessed 14 December 2023).
[20] Strobelt H, Webson A, Sanh V, et al. Interactive and visual prompt engineering for ad-hoc task adaptation with large language models. IEEE Trans Vis Comput Graph 2023;29:1146–56.
[21] Acar OA. Are your students ready for AI? Harvard Business Publishing; 2023. Available at: https://hbsp.harvard.edu/inspiring-minds/are-your-students-ready-for-ai (last accessed 14 December 2023).
[22] Weaver KF, Morales V, Dunn SL, Godde K, Weaver PF. An introduction to statistical analysis in research: with applications in the biological and life sciences. Wiley; 2017. Available at: https://onlinelibrary.wiley.com/doi/book/10.1002/9781119454205 (last accessed 26 February 2024).